Comparison of male and female HIV Seroprevalence rates from a coal mining community and mobile clinic in Mpumalanga, South Africa

Hitesh Hurkchand

A research report submitted to the Faculty of Health Sciences, University of the Witwatersrand, Johannesburg, in partial fulfillment of the requirements for the degree

Of

Master of Science in Medicine in the field of Epidemiology and Biostatistics

Johannesburg, 2007

DECLARATION

I, Hitesh Hurkchand declare that this research report is my own work. It is being submitted for the degree of Master of Science in Medicine in the field of Epidemiology and Biostatistics in the University of the Witwatersrand, Johannesburg. It has not been submitted before for any degree or examination at this or any other University.

_____

28th of November, 2007

In loving memory of my father, Pravinchundra Hurkchand

(1937 – 2004)

PUBLICATIONS AND PRESENTATIONS

Comparison of HIV seroprevalence between males and females at clinic and community level in Mpumalanga South Africa

H.P Hurkchand, J.B. Levin, H.Makuluma. (15th International AIDS Conference, Bangkok, Thailand, July 2004)

Measuring the impact of HIV and STIs in a cluster designed household prevalence survey in a coal mining community, eMbalenhle, Mpumalanga, South Africa.

H.P.Hurkchand, J.B.Levin, H.Makuluma, M.Molapo, N.Molefe. (15th International AIDS Conference Bangkok, Thailand, July 2004)

Measuring the impact of HIV and STIs at community mobile clinics in a coal mining town, Dunusa, Mpumalanga, South Africa.

H.P.Hurkchand, J.B.Levin, H.Makuluma, M.Molapo, N.Molefe. (15th International AIDS Conference Bangkok, Thailand, July 2004)

ABSTRACT

Abstract Title: Comparison of HIV seroprevalence between males and females at clinic and community level in Mpumalanga South Africa.

Background: Two cross-sectional surveys were conducted in Embalenhle community (February 2002) and Dunusa community mobile clinics (November 2001), to establish prevalence of HIV and STIs (*Chlamydia trachomatis* and *Neisseria gonorrhea*).

Methods: Multiple logistic regression models were fitted to the combined data from the two sites, to identify factors associated with HIV prevalence and also to check whether the effects were consistent over the two sites.

Results: HIV Prevalence was 33.5% (30.2%vs.35.9% in males and females respectively, p=0.124) at community site and 34.8% at clinic site (22.8%vs.47.4% in males and females respectively, p=0.001). The models show a significant site by sex interaction i.e. the effect of sex differs in the 2 sites (p=0.036). After adjusting for agegroup and *Neisseria gonorrhea,* predicted probabilities from the logistic regression model shows that the sex difference is much greater in community mobile clinics (23%vs.44.1% in males and females respectively) than at the community site (29.9%vs.34.9% in males and females respectively). After adjusting for site and *Neisseria gonorrhea,* the model showed an agegroup by sex interaction (p<0.001). Predicted probabilities show a difference, where HIV in males is higher than in females; in males in the 25-34 year age group from 18-24 years (36.3 vs 18.2 % respectively), while in females the prevalence is very similar in the 18-24 year and 25-34 year age groups. There were no interactions between *Neisseria gonorrhea* and other variables.

Conclusions: The different HIV–age distribution for males and females are consistent with the results of previous studies. We found that the sex difference in prevalence was much smaller at the community level than at the clinic level. The traditional interpretation of national antenatal surveillance data assumes a fairly large difference in male and female seroprevalence (a ratio of 7:10 is used in extrapolating results of the South African National antenatal seroprevalence survey to males). These results suggest that more work is needed in checking that assumption.

ACKNOWLEDGEMENTS

Dr Jonathan Levin for his absolute brilliance, understanding and endless patience in supervising this work.

The Council for Scientific and Industrial Research for permission to use the data.

Work colleagues at the Council for Scientific and Industrial Research for technical assistance and support.

TABLE OF CONTENTS

Page

LIST OF FIGURES

LIST OF TABLES

GLOSSARY OF ABBREVIATIONS

AIDS	-	Acquired Immune Deficiency Syndrome
ANC	-	Antenatal Clinic
ANRS	-	Agencé Nationale de Recherches sur le Sida
BV	-	Bacterial Vaginosis
CBOs	-	Community Based Organizations
d.f.	-	Degree of Freedom
DOH	-	Department of Health
DOTS	-	Directly Observed Treatment Short-Course
DSW	-	Department of Social Welfare
CADRE	-	Centre for AIDS Development, Research and Evaluation
CDC	-	Centre for Disease Control
CSIR	-	Council for Scientific and Industrial Research
DFA-TP	-	Direct Fluorescent Antibody Test Treponema Pallidum
EHO	-	Environmental Health Officer
FBOs	-	Faith Based Organizations
GCP	-	Good Clinical Practice
HASA	-	Hospice Association of South Africa
HBC	-	Home Based Care
HIV	-	Human Immune Deficiency Virus
HSRC	-	Human Sciences Research Council
HSV	-	Herpes Simplex Virus
ICHC	-	Integrated Community – based Home Care
MRC	-	Medical Research Council
NGOs	-	Non Governmental Organization
NSP	-	National Strategic Plan
PLWHA	-	People Living With HIV and AIDS

STIs	-	Sexually Transmitted Infections
TEBA	-	The Employment Bureau of Africa
TB	-	Tuberculosis
UNAIDS	-	United Nations – AIDS Program
USAID	-	United States Agency for International Development
WHO	-	World Health Organization

Key words and phrases

Household-survey, clinic-survey, HIV surveillance, HIV prevalence, South Africa, demographic impact, logistic regression modeling

1. Introduction

1.1 Statement of the problem and research question

Available data on HIV prevalence in sub-Saharan Africa suggest large differentials in male to female HIV seroprevalence in different population settings (1). This large differential suggests the need to further explore and quantify this gap in order to effect optimal planning in HIV prevention efforts. This study aimed to conduct an exploratory analysis of HIV seroprevalence data collected from two cross-sectional surveys to investigate the differences in sex-related HIV prevalence and the degree of extrapolation of HIV prevalence from females in a clinic setting to males in the general population; and also to investigate specific risk factors and whether or not these risk factors were consistent over the two sites.

1.2 Justification of the study

There has been an increasing need for estimates and projections of HIV prevalence in recent years for advocacy purposes, monitoring and evaluating trends of incidence, impact of relevant interventions and planning for future needs and resource allocations (2). A major assumption, using HIV serological testing, is that HIV prevalence found in antenatal populations can, with adjustment for the estimated male to female ratio, be used as a surrogate for HIV seroprevalence in the total 15-49 year population (2). This assumption, used in sub-Saharan Africa, is supported by limited community based HIV sero-surveys which suggest that HIV seroprevalence among antenatal females is a reasonable surrogate value for HIV seroprevalence in the general population and for the general male population (2). This assumption has not been validated. Measurement and/or estimation of the male to female ratio of HIV infections have been carried out using a variety of methods and assumptions. As many settings do not factor in a male to female ratio in their process of estimating their national

HIV seroprevalence, this could result in gross bias if antenatal data are used without any adjustment to estimate HIV prevalence in both males and females (2). All HIV prevalence estimations should try to ensure that the overall HIV prevalence estimated is consistent with the estimated male to female ratio. The study aims to explore the large differentials in male to female ratios.

1.3 Background and literature review

The HIV/AIDS epidemic globally is still on the increase. According to the UNAIDS 2005 global summary of the AIDS epidemic report, approximately 38.6 million (33.4 million - 46.0million) people were living with HIV by the end of 2005 (3, 4). About 4.1 million new HIV infections (3.4 million – 6.2 million) and an estimated 2.8 million deaths were recorded in the same year (3).

Overall the HIV incidence rate is believed to have peaked globally, although the rate is still increasing in several countries (4)

Globally, 3.8% (3.0%-4.7%, n=54 countries) of young females, aged 15-24 years are HIV infected (4).

Approximately 60% (25.8-40.3 million) of all people living with HIV in the world live in sub-Saharan Africa despite having just a meagre 10% of the world's population (4). This highlights the magnitude of the HIV/AIDS burden the sub-Saharan African Region is currently experiencing (4). The region had an estimated 3.2 million new infections and 2.4 million deaths due to AIDS in 2005 (4). Unfortunately, there is no convincing evidence yet of a decline in epidemic prevalence rates in southern Africa (4).

The available data on HIV prevalence in sub-Saharan Africa indicate substantial heterogeneity in the spread of HIV across the continent. Large differences exist

between and within countries and a range of biological, behavioural and contextual arguments have been advanced to explain these differences (1, 5).

Biological explanations include those that focus on different sub-types of HIV-1 and variation in prevalence of other sexually transmitted infections (STIs). Male circumcision, inextricably linked with cultural practices, is also considered an important biological factor in understanding the differentials in risk of HIV transmission and prevention (6, 7). Differences in sexual practice, whether or not in relation to the response to the epidemic, have also been considered as important explanatory factors, as have differences in underlying factors, such as mobility, infrastructure and poverty.

An important attempt to assess factors affecting the differential spread of HIV in urban Africa was made in a multi-centre study of four cities (8, 9). Two cities had a very high prevalence of HIV in the general population (Kisumu, Kenya and Ndola, Zambia) and two cities had much lower prevalence (Yaounde, Cameroon and Cotonou, Benin). The study design combined ecological comparisons across populations with individual level analysis within populations, and it was concluded that differences affecting the efficiency of HIV transmission, notably lack of male circumcision and prevalence of ulcerative STIs, including Herpes Simplex Virus-2 (HSV-2), were the most important explanatory factors, while differences in high risk sexual behaviour appeared to play a much smaller role. The main differences in the latter pertained to earlier sexual debut, earlier marriage and larger age difference between spouses in the high prevalence cities (8).

Boisson et. al. propose a model that uses HIV prevalence estimated for pregnant women from unlinked anonymous surveys to determine the prevalence of HIV in women in the same population (10). The model assumes the ratio of prevalence in pregnant women to that in all women is influenced by HIV-related risk behaviours that are different for pregnant and non-pregnant women and also by differences in fertility level among infected and uninfected women. The ratio is

affected by biases, likely to be culturally or socially specific, and which are qualified and quantified by the model.

Like most countries in the sub-Saharan region, South Africa is also bearing the brunt of the HIV/AIDS pandemic with a reported 5.6 million people living with the HI virus giving a prevalence rate of 11.4% (11). Of these, 15.6% are adults in the economically active age groups, and 15.2% of persons in the 15-49 age group were HIV positive. The high prevalence of HIV/AIDS impacts negatively on productivity and sustainability of all developmental programs due to significant reduction of human resource capacity in all sectors.

A study done by the South African Department of Health involving more than 16 500 pregnant women attending antenatal services across all the nine provinces in 2005 found that 30.2% (95% CI 29.1; 31.2) were HIV positive (12). Amongst all the provinces, Kwa-Zulu Natal province recorded the highest HIV rate of 39.1% (95% CI 36.8; 41.4) (11). The most recent HIV antenatal survey estimated a national sero-prevalence of 29.1% (95% CI 28.3; 29.9) (13).
Data from ANC surveys for 2002, 2003 and 2004 were analyzed to establish HIV prevalence trends that showed an increase of HIV prevalence in HIV positive pregnant women from 27.9(95% CI 26.8; 28.9) in 2003 to 30.2% (95% CI 29.1; 31.2) in 2005 (12). After adjusting for the effect of province and age group, a three year increase was statistically significant between the period 2002 and 2004 ($p<0.001$) (14).

In many, if not all settings, sentinel surveillance in antenatal clinics (ANC) is the chief source of routine HIV prevalence information, but provides no information on males (14).

The Human Sciences Research Council (HSRC), in partnership with the Medical Research Council (MRC), Centre for AIDS Development, Research and

Evaluation (CADRE), and Agencé Nationale de Recherches sur le Sida (ANRS), conducted South Africa's first national household study of HIV/AIDS in 2002 (11). The survey included gathering of data on HIV prevalence, behaviour and communication. The national community household survey conducted in 2002 (11), demonstrated an HIV prevalence in males 12.8%, and females 17.7%. However the overall response rate (at the level of having an HIV test) was low and the results may have been subject to non-response bias. The result of this survey found an HIV seroprevalence of 17.7% in females aged 15-49 years old and show a quite a large difference to the 2001 Department of Health Antenatal survey which found an HIV prevalence of 24.8% in Antenatal Clinic Attendees (15).

The 2005 SA National HIV survey was the second household survey conducted by the Human Sciences Research Council and sampled nearly twice as many study participants as the first (16). The results presented below are similar to those of the 2002 HSRC survey, however they are quite different from the SA National Department of annual Antenatal Clinic Health Survey of pregnant women and also disagree with UNAIDS estimate of 21.5% of South Africans aged between 15 -49 years living with HIV at the end of 2003.

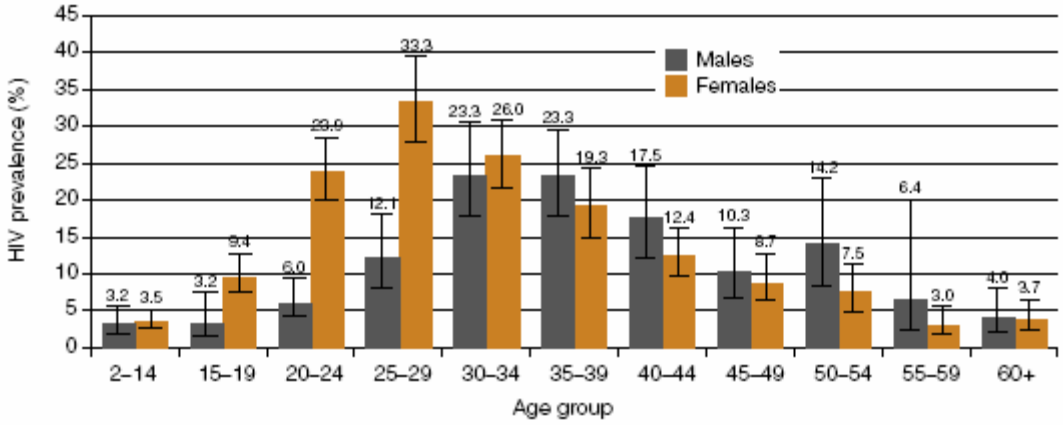

Figure 1 Distribution of HIV seroprevalence between males and females

Figure 1 illustrates the distribution of HIV seroprevalence between males and females (HSRC, 2005) with the females having the highest rate in the 15-34 year age band.

Table 1. Distribution of HIV seroprevalence among females compared between the National Department Antenatal Survey and HSRC Survey

Age group	Females 2005 n=5650		Pregnant in the last 24 months n=918		Antenatal Survey 2004 n=15 976	
	% (n)	95% CI	% (n)	95% CI	% (n)	95% CI
15-19	9.5 (1 153)	7.1-12.4	19.7 (79)	10.0-35.2	16.1 (3130)	14.7-17.5
20-24	23.9 (1 182)	19.8-28.4	25.0 (303)	18.0-33.7	30.8 (4991)	29.3-32.3
25-29	33.3 (598)	27.7-39.4	32.1 (184)	21.6-44.7	38.5 (3702)	36.8-40.3
30-34	26.0 (691)	21.5-30.9	20.6 (157)	12.9-31.1	34.4 (2510)	32.2-36.6
35-39	19.3 (727)	14.9-24.6	15.7 (126)	9.4-25.3	24.5 (1 261)	21.9-27.2
40-49	10.7 (1 299)	8.6-13.3	11.3 (69)	4.7-24.8	17.5 (382)	14.0-21.0
Total	20.2	18.3-22.2	23.2	19.0-28.1	29.5	28.5-30.5

Table 1 compares HIV prevalence among females in the 15–49 year age group with findings of the annual antenatal survey conducted by the Department of Health in 2004. HIV prevalence in five-year age bands for all females aged 15–49, and for those who were pregnant in the last 24 months in the 2nd HSRC Community Based Survey is provided for comparison (16). The overall HIV prevalence in females participating in the 2005 household survey was 20.2% (16). In the survey sample of females who were pregnant in the last 24 months (n = 918), 23.2% were HIV positive (16). These figures are lower than the 29.5% HIV prevalence found in the 2004 antenatal survey (14). However, the household survey included females of all race groups, regardless of whether or not they were sexually active, whilst the antenatal survey is only representative of pregnant females using government clinics. Taking into account a differential utilization rate of these clinics by race and income group well over 90% of the females in the 2004 antenatal survey were African females (14).

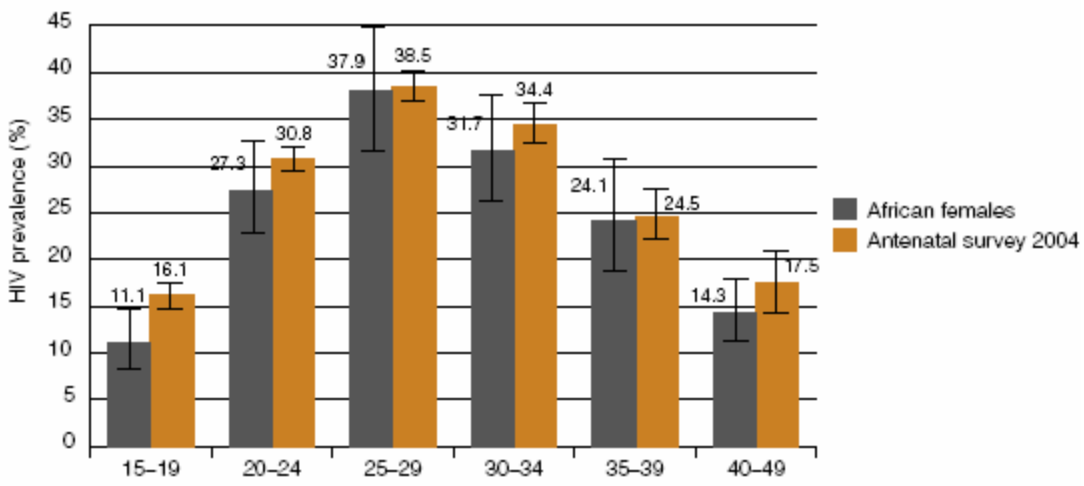

Figure 2 Distribution of HIV seroprevalence by agegroup between females in the National Department of Health Survey and HSRC Survey

Figure 2 shows a descriptive comparison of African females from both HSRC Community Based Survey (2005) and the National Department of Health ANC survey (2004). Although a descriptive comparison, both surveys predict similar overall levels of HIV prevalence, and show that the growing epidemic is a cause for concern both in the general population and the targeted ANC population (11, 14).

Data from the first HSRC study however suggests a somewhat different provincial prevalence picture (11). Gauteng, Free State and Mpumalanga had the highest prevalence rates, while all other provinces had prevalence rates that were about or below 10%. KwaZulu Natal ranked fourth and the Eastern Cape has the lowest prevalence. The observed prevalence for women aged 15-49 years old in the Western Cape of 18.5% was much higher than that observed from antenatal data. This was the only province, as described in the HSRC community based survey, where the HIV prevalence derived from the household survey was much higher than that derived from the antenatal data.

The traditional interpretation of national antenatal surveillance data assumes a fairly large difference in male and female seroprevalence (a ratio of 7:10 is used in extrapolating results of the South African National antenatal seroprevalence survey to males) (17).

Zaba et. al. (18) propose that data on age of the father be collected in antenatal sentinel HIV sero-prevalence surveys to estimate overall and age-specific male HIV prevalence; however this method may produce an under-estimate of male prevalence, especially in the oldest age groups, because of the fertility reducing effects of HIV and the age pattern of discordancy. Other anticipated drawbacks are: 1. reliance on questionable assumptions (e.g. that both partners share the same sero-status); 2. exclusion of men who are without regular partners or not currently sexually active or infertile or contraceptive users; 3. willingness and ability of women to disclose information.

1.4 Study Objectives

1.4.1 Aim

To conduct secondary data analysis to look at HIV prevalence across the two sites and to examine factors associated with HIV.

1.4.2 Objectives

1. To compare the prevalence of HIV and selected STIs between the two sites.

2. To look at the relative risk of HIV infection for females versus males and compare this between the two sites.

3. To look for specific risk factors for HIV infection and whether these are consistent over the two sites.

2. Methodology

2.1 Study design

This study was a secondary analysis and made use of data that was originally collected in a cross-sectional study.

2.2 Description of the study population, measurements taken and instruments used

The original study commissioned by the Council for Scientific and Industrial Research, used a cross sectional study design (19). The aim of this study was to determine HIV prevalence in Embalenhle community in November 2001 and Dunusa Mobile clinics in February 2002. Embalenhle and Dunusa are coal mining towns in Mpumalanga and major labour sending areas.

The study population comprised of adults (males and females) who were included into the study from both community and clinic for Embalenhle and Dunusa respectively. All males and females aged 18 years or older who gave verbal consent and who agreed to provide a specimen for an unlinked anonymous HIV test were included in the study. This study is a secondary exploratory analysis. Sites were selected according to the initial study design that was conducted by the CSIR (described below).

2.2.1 Description of original sampling for Embalenhle

For purposes of the cross sectional study considered in this secondary analysis, participants from Embalenhle were selected using a two stage household sample design. Field workers recruited participants from the community based on a representative sample and patients presenting in the community (Embalenhle).

The rational for site selection was to describe HIV seroprevalence in these settings so as to implement HIV interventions in these settings. The sites were

selected conveniently in the initial operational study. Embalenhle was mapped into 3 sections which served as clusters. HIV testing was unlinked and anonymous. Study participants were recruited from sections A, B and C and tested at Zone 14 and Zone 4 Clinics in Embalenhle.

2.2.2 Description of original sampling for Dunusa

Participants in Dunusa mobile clinics were selected using systematic sampling. The rational for site selection was to describe HIV seroprevalence in these settings so as to implement HIV interventions in these settings. The sites were selected conveniently in the initial operational study. Field workers recruited participants from the community based on a representative sample and patients presenting at clinics (Dunusa). The majority of patients (greater than 90%) at Dunusa were clinic attendees. Dunusa mobile clinic and Impungwe Hospital participated in the survey.

2.3 Data management, processing and data analysis

2.3.1 Data variables

The study made use of biological markers. In addition, a socio-demographic and behavioural questionnaire was administered to participants during study visits.

The variables available for the secondary data analysis are presented below. The analysis was constrained to measurements of the variables below and the investigator was not allowed to use socio-demographic variables that was also captured.

1. Age
2. Sex
3. HIV status
4. STIs

    a. *Neisseria gonorrhoeae*
    b. *Chlamydia trachomatis*
5. Site
    a. Dunusa
    b. Embalenhle

CSIR has access to these variables only. The data from the socio-demographic questionnaire would have added a deeper insight into investigating the relationship between sexes at the two sites. These variables were however not available from the service provider due to contractual reasons.

2.3.2 Data cleaning

Some of the problems that the study encountered with the two data sets were:

- Data was captured into a Microsoft excel file and double data entry was not performed.
- Missing values were not appropriately coded
- The service providers for the original study collected data on participants who did not meet the eligibility criteria for the survey. The criteria for inclusion included age specific criteria of > 18 years and less than 50 years for inclusion into the analysis.

2.3.3 Database cleaning

All data cleaning was performed in STATA 8. Frequency checks were performed on all variables to determine the extent of missing values in the database and missing data was encoded using standard operating procedures. These procedures included replacing all missing data with specific numerical values of

‘99’ and ’999’ which was subsequently excluded from the basic and multivariate analysis.

2.3.4 Data coding

In order to facilitate data analysis, most raw data variables required further manipulation and/or grouping resulting in the creation of new database variables. Data was exported from MS Excel to STATA 8.0, via StatTransfer 7.0, to facilitate higher level coding and manipulation of data. Several secondary variables were created to describe further associations between categorical variables and HIV status; certain variables were recoded as categorical variables, many with only two levels to determine the level of risk exposure. Please refer to the analysis section for details on coding of the results.

2.3.5 Data archiving

Each participant has a confidentially-named clinic record, containing details of visit date, signed consent and laboratory results. The records were identified by a bar-code and unique study number only. Copies are kept at the CSIR Miningtek.

Individuals were given an information sheet. The records and logs were kept in a secure location at the managing centre, for the duration of the study.

2.3.6 Data ownership

Each case record has a confidentially-named record, containing details of interview dates and data collected. It was the responsibility of the investigators to ensure that forms were adequately and fully completed. The records are kept by the investigator to ensure confidentiality.

All questionnaires were reviewed by the investigator and the team, to ensure that they were accurate and complete. Data was entered onto a computerized database, and preparation of this and the data files for analysis was conducted by the investigator and the team.

The data generated in this study belongs to investigators and the Powerbelt Steering Committee.

2.4 Statistical considerations

2.4.1 Sample size

All available data from the original survey was used in this survey. 690 study participants (399 female and 291 male) in Embalenhle were sampled for the survey; and 155 study participants (76 female and 79 male) sampled in Dunusa. There is a difference in the sample sizes in the two sites and this may be attributed to the fact that there was a smaller population at the Dunusa clinics, as well as budgetary constraints of the project.

2.4.2 Analysis

Analysis was conducted using STATA version 8.

2.4.3 Descriptive analysis

1. Descriptive frequencies of the distribution of HIV, sex, age, STI, any STI and site
2. Students t-test for significant difference in mean age of males and females

3. Chi square tests for associations between HIV status and categorical variables (such as sex, individual STIs, any STI and site). In the case of ordinal explanatory variables (such as agegroup) a chi square test for trend was used.
    a. We created a variable ‘Any STI’ and determined the seroprevalence for participants having any STI in both communities.
4. By means of inferential statistics, confidence intervals were used to show precision with which prevalences are estimated in different sub-groups. Confidence intervals were also used for odds ratios. The process for calculating this interval is complicated as this ratio is not normally distributed. The interval contains the value of the true odds ratio with 95% confidence. If the odds are the same in each group of comparison, then the value of the odds ratio is approximately 1. If the confidence interval for the odds ratio does not contain 1, then there is strong evidence of association between exposure and the risk of infection, while if the confidence interval contains 1, then we have insufficient evidence to conclude that there is an association.

2.4.4 Analysis methods for logistic regression models

1. Logistic regression models were fitted.
2. Factors investigated were age, sex, site and STI presence, as well as pairwise interactions between these factors.
3. Model selection was conducted using Likelihood-Ratio tests.
4. Model interpretation: In order to interpret interactions (effect modification), we adjusted for variables in the model and observe predicted probabilities.

2.4.4.1 Maximum likelihood estimation (20, 21)

In linear regression, the method of least squares is used to estimate regression coefficients, i.e. we choose those estimates of α and β that minimize the sum of squared residuals (20, 21). This approach does not work well in logistic regression, or for the entire family of generalized linear models. Instead we use another approach called maximum likelihood estimation (20, 21).

The maximum likelihood estimate of π is the value of π that assigns the greatest probability to the observed outcome (20, 21). In general, maximum likelihood estimates do not have simple closed solutions but must be found iteratively using numerical methods (20, 21).

A likelihood function looks deceptively like a probability density function. It is important to realize that they are different (20, 21). A probability density function uses fixed values of the model parameters and indicates the probability of different outcomes under this model (20, 21). A likelihood function holds the observed values fixed and shows the probability of this outcome for the different possible values of the parameters (20, 21).

We discuss how the likelihood approach can be used to provide a general means of hypothesis testing. A hypothesis test is based on calculating a test statistic and its corresponding P-value (also known as the significance level), in order to assess the strength of the evidence against the null hypothesis (of no association between exposure and outcome in the population) the smaller the P Value, the stronger the evidence against the null hypothesis (22).

There are 3 different types of tests based on the log likelihood:

1. The Likelihood Ratio test, based on the value of the log likelihood ratio at the null value of the parameter (22).
2. The Wald test uses the value of the fitted quadratic approximation to the log likelihood ratio at the null, rather than the actual value of the log likelihood ratio at this point (22).
3. The Score test, based on fitting an alternative quadratic approximation to the log likelihood ratio which has the same gradient and curvature at the null value of the parameter, rather than at the maximum likelihood estimation (22).

2.4.4.2 Choice of test for this project (20, 21, 22)

Likelihood ratio test in regression models

Hypothesis testing in regression models can be carried out using either Wald tests or likelihood ratio tests (22). Likelihood ratio tests tend to be favoured for all but the simplest of cases, for the following reasons:

1. the lack of dependence of the likelihood ratio statistic on the scale used for the parameter(s) of interest
2. the ease with which the calculation and interpretation of likelihood ratio statistics can be carried out in a more complex situations
3. In contrast, although the Wald tests are directly interpretable for exposure variables which are represented by a single parameter in the regression model, they are less useful for a categorical variable which is represented by a series of indicator variables in the regression model

The likelihood ratio test described above for a single exposure is a special case of a more general likelihood ratio test that applies to more complex situations involving several model parameters (22). An example is in the regression

modelling where we have estimated the effect of a categorical exposure variable using *k* indicator variables and wish to test the null hypothesis that the exposure has no association with the outcome. In such situations we wish to test the joint null hypothesis that k parameters equal their null values. The likelihood ratio test is based on comparing the log likelihoods obtained from fitting the following 2 models:

1. L_{exc}, the log likelihood of the model excluding the parameters to be tested
2. L_{inc}, the log likelihood of the model including the parameters to be tested

Then the likelihood ratio statistic (LRS) has a χ^2 distribution with degrees of freedom equal to the number of parameters omitted from the model (22).

The three tests above generalize to more complicated situations. Given a sufficiently large sample size, all of these methods are equivalent (20, 21). However likelihood ratio tests and score tests are more accurate than the Wald test for most problems encountered in practice. The likelihood and score tests are used in practice for this reason (20, 21).

The likelihood ratio test has the property that is unaffected by transformations of the parameter of interest and is preferred over the score test for this reason.
The Wald test is much easier to calculate than the other two, which are often not given by statistical software packages. It is common practice to use the Wald test when it is the only one that can be easily calculated.

Wide divergence between these three tests can result when the log likelihood function is poorly approximated by a quadratic curve. In this case, it is desirable to transform the parameter in such a way as to give the log likelihood function a more quadratic shape.

When interpreting the effects of factors in the presence of interactions, we have to remember (a) that odds ratios are multiplicative rather than additive and (b) a constant factor change in the odds does not correspond to a constant change or a constant factor change in the probability. For this reason interpretations based on predicted probabilities are especially useful in jointly examining the effect of two categorical explanatory variables (holding other variables in the model constant) by means of a table of predicted probabilities (23).

For the proposed set of data analyses, we made use of the likelihood ratio test and investigated changes in deviance as a means of selecting a best model for our data. Deviance is defined as a quality of fit statistic for a model (similar to R-squared for ordinary least squares) that is often used for statistics hypothesis testing. The expression of deviance is simply – 2 times the log likelihood of the model of fit. While the deviance has a derivable asymptotic distribution under the assumption of the correctness of the model, it is rarely used. Instead the deviance is used to compare 2 models in particular of generalized linear models where it has a similar role to residual variance from analysis of variance in linear models. Suppose in the framework of the generalized linear model, we have two nested models, M_1 and M_2. In particular, suppose that M_1 contains all of the parameters in M_2, and k additional parameters. Then, under the null hypothesis that M_2 is the true model, the difference between the deviances for the two models follows an approximate chi-squared distribution with k-degrees of freedom.

2.5 Ethical considerations for the original baseline survey

The committee for ethical clearance of the University of the Witwatersrand approved standard subject information sheets (reference No. 001109), that were translated into English, Sotho and Zulu and distributed in the community and clinics. The study did not make use of formal written informed consent as HIV testing was unlinked and anonymous.

A protocol for feedback of STI results to the relevant clinic authorities and community notification was set up. It was agreed by clinics and hospitals of Evander Hospital in Embalenhle that health authorities would provide treatment for participants testing positive for sexually transmitted infections. Urine was collected for the detection of sexually transmitted infections *N. gonorrhoeae* and *Chlamydia trachomatis* by polymerase chain reaction. All urine specimens were bar-coded. Patients were given a copy of bar codes if they decided to return for results and treatment. Urine specimens were collected and transported by Contract Laboratory Services daily to Johannesburg.

Men and Women were given health education about STIs including HIV prevention messages at each visit so that they had an understanding of the following:

- transmission,
- symptoms and signs,
- impact on their health, their partner's health and their children's health,
- treatment and how to access it, and
- how to protect themselves, their sexual partners and their children from future infections
- HIV/STI counselling

Participants were informed about the purpose of the study, the commitment of time, and their saliva specimen collection for HIV testing. They were informed that these procedures were undertaken throughout the study to ensure that volunteers clearly understood that there may be reasons for ineligibility to participate in the study. They were told that they were free to withdraw at any time with no jeopardy to their medical care with respect to treatment of STI infection. They were told that screening for sexually transmitted infections other than HIV would also be performed during the study and that this would be done

by collection of urine specimens. They were advised to discuss the study with their regular sexual partner but the consent of their partner would not be essential. Written informed consent was not administered by the study staff at screening and enrolment.

Participants were counselled about the importance of condom use at the study visit and this was an active procedure undertaken by an experienced counsellor. Hypo-allergenic non-spermicidal lubricated condoms were given to the participants.

HIV testing was unlinked and anonymous. Participants were offered the option of whether to participate or not in the prevalence study but they were not to receive their HIV results.

HIV testing was conducted using the ORASURE Saliva test kit. Age and sex were the only information collected with the specimens. No personal identifiers were collected. Specimens were stored and transported by Contract Laboratory Services daily to Johannesburg. There was no HIV pre- and post-test counselling offered. Participants were not given the option of receiving their HIV test results.

2.5.1 Confidentiality, ethics and responsibilities of the investigators

Full medical and social confidentiality is maintained with all data records. No names were displayed on any form of report or publication and interview assessment forms were kept by the investigator of the study to ensure confidentiality.

The investigators were responsible for obtaining ethical approval for the study (ethical clearance was obtained by the HIV Management Solution, Wits Health Consortium). The investigators were responsible for the implementation and

optimization of the standard operating procedures to be used in the semi-structured interviews.

The investigators, based at the CSIR, were responsible for the reports of the progress of the study and financial reports.

The study was conducted according to ICH-GCP guidelines and the current version of the Declaration of Helsinki (South Africa 1996.). It was the responsibility of the local investigators to abide by this.

2.5.2 Ethical considerations

The study made use of secondary data analysis from an original study, which attained ethical clearance from the Committee for Ethical Clearance, University of the Witwatersrand.

3. Results

Six hundred and ninety participants from Embalenhle and 155 participants from Dunusa were included in the analysis. The sex distribution for Embalenhle was 399 females and 291 males, and for Dunusa 76 females and 79 males. There was a significant difference between the distribution of sex across sites (p=0.04), table 2.

Table 2. Descriptive statistics for male and female study participants by site in an HIV sero-prevalence study in Mpumalanga in 2003

	Embalenhle		Dunusa		P Value
	n	%	n	%	
Sex					**0.04***
Male	291	42.17	79	50.97	
Female	399	57.83	76	49.03	
Agegroup (yrs)					**0.37***
17-24	192	27.83	37	23.87	
25-34	255	36.96	69	44.52	
35-44	192	27.83	38	24.52	
45-50	51	7.39	11	7.10	
STI Prevalence					
HIV					**0.75***
Pos	231	33.48	54	34.84	
Neg	459	66.52	101	65.16	
N. gonorrhoeae					**0.23***
Pos	29	4.35	10	6.67	
Neg	638	95.65	140	93.33	
C. trachomatis					**0.22***
Pos	48	7.25	15	10.27	
Neg	614	92.75	131	89.73	
Any STI					**0.28***
Pos	70	10.57	20	13.7	
Neg	592	89.43	126	86.3	

*Chi-square test

There also seemed to be a symmetrical distribution of sex by agegroup with majority of the male participants presenting between the ages 25 and 34 years with females showing a similar pattern with majority of the women presenting also between the ages 25 and 35 years. The mean age for males at Embalenhle was 31.75 years (95% CI, 30.73; 32.70) and 31.34 years in Dunusa (95% CI, 30.73; 32.70), p=0.39. The mean age for females at Embalenhle was 31.01 years (95% CI, 30.17; 31.84) and 31.89 years in Dunusa (95% CI, 30.19; 33.59), p=0.71.

Table 2 also describes the prevalence of STIs across sites as follows: HIV seroprevalence in Embalenhle was 33.48% (95% CI, 29.96; 37.13) and 34.84% (95% CI, 27.37; 42.89) in Dunusa, p=0.75. *N. gonorrhoea* seroprevalence in Embalenhle was 4.45% (95% CI, 2.93; 6.18) and 6.67% in Dunusa (95% CI, 3.24; 11.92), p=0.23. *C. trachomatis* seroprevalence in Embalenhle was 7.25% (95% CI, 5.39; 9.49) and 10.27% in Dunusa (95% CI, 5.87; 16.39), p=0.22. The seroprevalence of having any STI in Embalenhle was 10.57% (95% CI, 8.33; 13.17) and 13.7% in Dunusa (95% CI, 8.57; 20.36), p=0.28.

Table 3a describes the distribution of independent variables by sex and the seroprevalence of STIs by sex as follows: HIV seroprevalence in males was 28.65% (95% CI, 24.09; 33.55) and 37.68% in females (95% CI, 33.31; 42.21), p=0.006. *N. gonorrhoeae* seroprevalence in males was 3.88% (95% CI, 2.14; 6.42) and 5.48% (95% CI, 3.58; 7.99) in females, p=0.29. *C. trachomatis* seroprevalence in males was 8.15% (95% CI, 5.52; 11.49) and 7.52% in females (95% CI, 5.27; 10.35), p=0.75. The seroprevalence for having any STI in males was 10.39% (95% CI, 7.03; 14.04) and 11.73% in females (95% CI, 8.91; 15.06), p=0.55.

Table 3a. Descriptive statistics for participants by sex in an HIV sero-prevalence study in Mpumalanga in 2003

	Male		Female		P Value
	n	%	n	%	
Site					**0.04***
Embalenhle	291	78.65	399	84	
Dunusa	79	21.35	76	16	
Agegroup (yrs)					**0.53***
17-24	104	28.11	125	26.32	
25-34	133	35.95	191	40.21	
35-44	102	27.57	128	26.95	
45-50	31	8.38	31	6.53	
STI Prevalence					
HIV					**0.006***
Neg	264	71.35	296	62.32	
Pos	106	28.65	179	37.68	
N. gonorrhoeae					**0.29***
Neg	347	96.12	431	94.52	
Pos	14	3.88	25	5.48	
C. trachomatis					**0.74***
Neg	327	91.85	418	92.48	
Pos	29	8.15	34	7.52	
Any STI					**0.55***
Neg	319	89.61	399	88.27	
Pos	37	10.39	53	11.73	

*Chi-square test

Table 3b. Distribution of Age by Site and Sex

			Males		Females		
			n=291		n=399		
Embalenhle	**Mean**	**95% CI**	**Mean**	**95% CI**	**Mean**	**95% CI**	**P Value**
Age	31.32	30.67; 31.97	31.75	30.73; 32.77	31	30.17; 31.84	0.26

			Males		Females		
			n=79		n=76		
Dunusa	**Mean**	**95% CI**	**Mean**	**95% CI**	**Mean**	**95% CI**	**P Value**
Age	31.61	30.35; 32.88	31.34	29.44; 33.24	31.89	30.19; 33.59	0.67

*Two-sample test with equal variances

Table 4. Comparison between male and female study participants by HIV serostatus in an HIV sero-prevalence study in Mpumalanga in 2003

	HIV Pos		HIV Neg				
	n	%	n	%	Unadjusted OR	95% CI	P Value
Site							
Embalenhle	231	33.48	459	66.52	1	reference	
Dunusa	54	34.84	101	65.16	1.06	0.73-1.53	**0.75***
Sex							
Male	106	28.65	264	71.35	1	reference	
Female	179	37.68	296	62.32	1.51	1.12-2.01	**0.006***
Agegroup (yrs)							
17-24	74	32.31	155	67.69	1	reference	
25-34	135	41.67	189	58.33	1.5	1.05-2.13	**0.03***
35-44	62	26.96	168	73.04	0.77	0.52-1.56	**0.21***
45-50	14	22.58	48	77.42	0.61	0.32-1.18	**0.14***
STI Prevalence							
N. gonorrhoeae							
Neg	255	32.78	523	67.22	1	reference	
Pos	19	48.72	20	51.28	1.95	1.01-3.72	**0.04***
C. trachomatis							
Neg	241	32.35	504	67.65	1	reference	
Pos	30	47.62	33	52.38	1.9	1.13-3.19	**0.02***
Any STI							
Neg	231	32.17	487	67.83	1	reference	
Pos	41	45.56	49	54.44	1.76	1.12-2.75*	**0.01**

*Chi-square test

Table 4 describes the distribution of independent variables by HIV sero-status. There was no evidence that HIV prevalence differed across sites {(OR = 1.06, 95% CI (0.73; 1.53 P=0.75)}. HIV serostatus differed across agegroup with the highest prevalence in the agegroup 25-34 years, p=0.03. Participants presenting positive with *N. gonorrhoeae* were more likely to be HIV positive {OR=1.95, (95% CI, 1.01; 3.72, p=0.04)}, 1.95 times more likely to be HIV positive when presenting with *C. trachomatis* {(95% CI, 1.13; 3.19, p=0.02)} and 1.76 times more likely to be HIV positive when presenting with any STI {(95% CI, 1.12; 2.75, p=0.01)}.

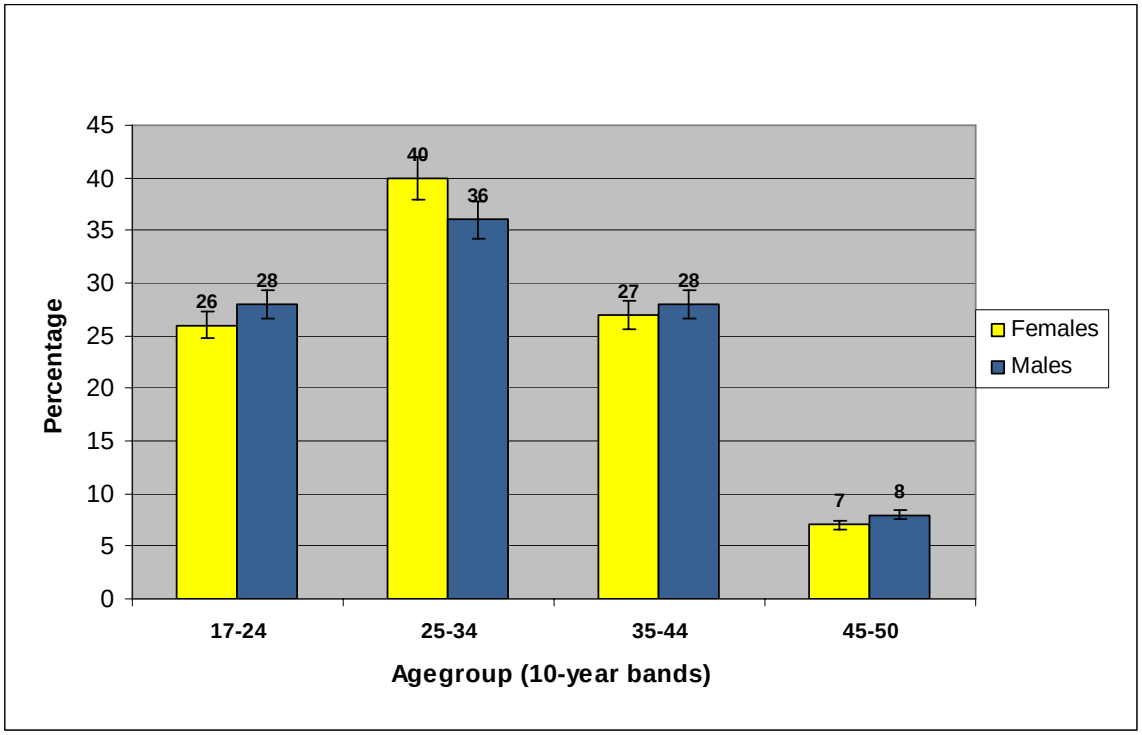

Figure 3 Distribution of agegroup by sex

Figure 3 illustrates the distribution of agegroup by sex with the majority of females presenting between the ages of 25 and 34 years and peaking in that same agegroup. The distribution of males was somewhat more uniform across age groups.

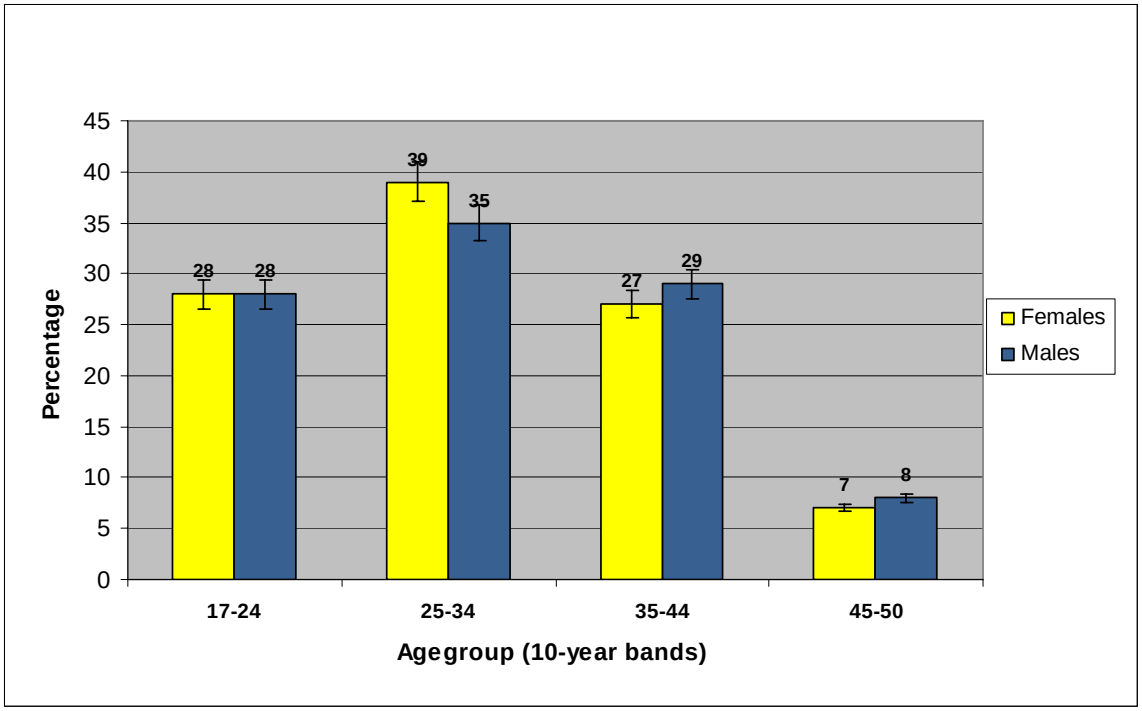

Figure 4 Distribution of agegroup by sex - Embalenhle

Figure 4 illustrates the distribution of agegroup by sex in the Embalenhle site with the majority of the females presenting between the ages of 25 and 34 years and peaking in the same agegroup. The distribution of males is unimodal with a peak in the 25-34 year agegroup; the frequency of males in 35-44 year agegroup was higher than in the 17-24 year agegroup.

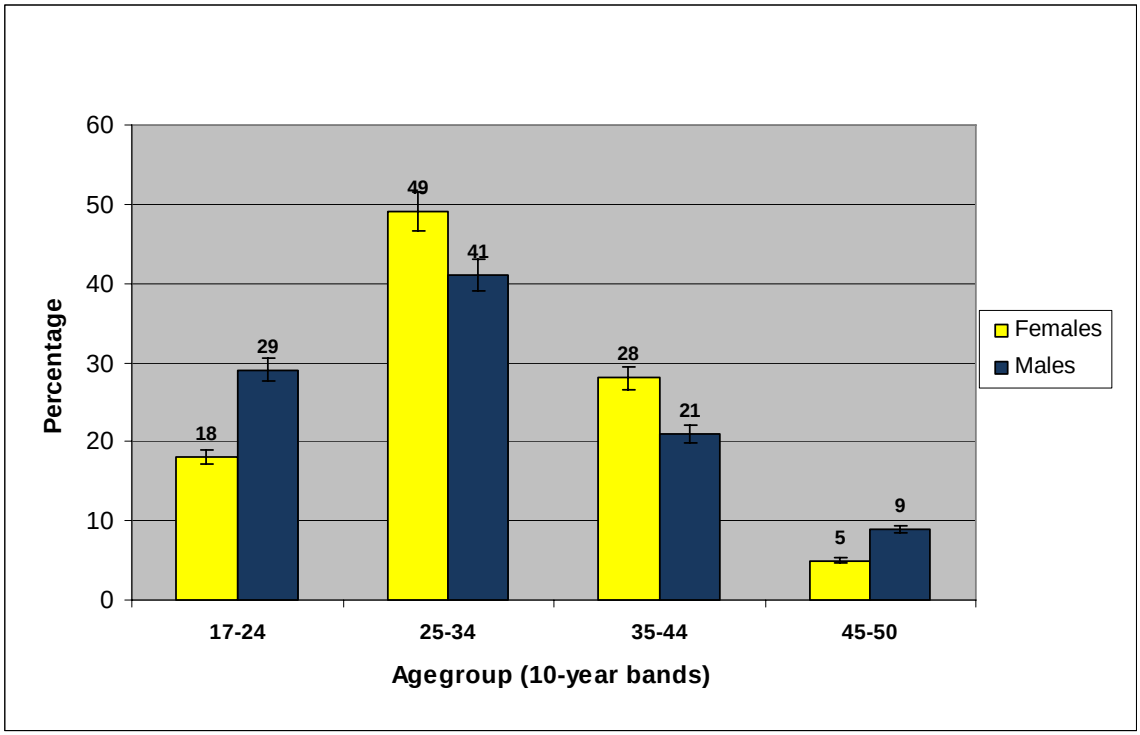

Figure 5 Distribution of agegroup by sex - Dunusa

Figure 5 illustrates the distribution of age by sex in Dunusa and shows a similar pattern to Embalenhle for males and females. The peak for females is still in the 25-34 year agegroup and the distribution for males is unimodal with a peak in the 25-34 year agegroup.

3.1 Model Selection

The analysis for this study involved logistic regression modelling and model comparisons. For purposes of the analysis, we describe the output of 3 logistic regression models and the model comparisons as highlighted in section 2.4.4. All models used HIV as the dependant variable and we present predictors for HIV in the outputs. Table 5 describes the candidate models, logistic regression outputs including the variables and interactions used for model comparisons.

Table 5. List of All Candidate Models

Variables		Model 1 (df)	Model 2 (df)	Model 3 (df)	Model 4 (df)	Model 5 (df)	Model 6 (df)
Sex		√ (1)	(1)	(1)	(1)	(1)	(1)
Site		(1)	(1)	(1)	(1)	(1)	(1)
Agegroup							
	18-24	√	√	√	√	√	√
	25-34	√ (1)	√ (1)	√ (1)	√ (1)	√ (1)	√ (1)
	35-44	√ (1)	√ (1)	√ (1)	√ (1)	√ (1)	√ (1)
	45-50	√ (1)	√ (1)			√ (1)	√ (1)
Chlamydia trachomatis		√ (1)	√ (1)	√ (1)			
Neisseria gonorrhoeae		√ (1)	√ (1)	(1)			√ (1)
Site*Sex		√ (1)	√ (1)	√ (1)	√ (1)	√ (1)	√ (1)
Sex*Agegroup		√ (3)	√ (3)	√ (2)	√ (2)	√ (3)	√ (3)
Sex**C.trachomatis*		√ (1)					
Sex**N.gonorrhoeae*		√ (1)	√ (1)	√ (1)			
Degrees of Freedom		13	12	10	7	9	10
Number of Observations		805	805	747	783	845	805
Exclusions		*Missing values include STIs, Chlamydia trachomatis and Neisseria gonorrhoeae*					
Log Likliehood		-489.27644	-489.2805	-460.42456	-483.94022	-516.54848	-492.12893
Goodness of Fit (P Value)		0.67	0.71	0.73	0.84	0.79	0.33

For the model comparison, we made use of the likelihood ratio test and investigated changes in deviance as a means of selecting a best model for our data. Please refer to 2.4.4.5 for model justification.

Table 6. Comparison of Model 1 and 2

	Model 1	Model 2	Difference in deviance
			Chi^2 - 2[-Log LiklihoodM2-(-Log Liklihood M1)]
Number of observations	805	805	
LR Chi^2	48.53	48.52	
df	13	12	
Prob > Chi^2	<0.001	<0.001	
Pseudo R^2	0.05	0.05	
Log Liklihood	-489.2764	-489.281	0.00812

Logistic Model for HIV, goodness of fit test

	Model 1	Model 2
Number of observations	805	805
Number of covariate patterns	50	50
Pearson Chi^2	31.79	31.81
Prob > Chi^2	0.67	0.71

Liklihood ratio test

Assumption	**Model 2 is nested in Model 1**
LR Chi^2	0.01
df	1
Prob > Chi^2	0.92

Table 6 illustrates the comparison of model 1 and 2. The change in deviance between model and 1 and 2 is 0.00812 on 1 d.f. Thus there is no evidence (p=0.92) that model 1 gives a better explanation of the data than model 2, so we can keep model 2 and do not need to use the more complex model 1. We can also note that there is no evidence of a lack of fit for either model.

Table 7. Comparison of Model 3 and 4			
	Model 3	**Model 4**	**Difference in deviance**
			Chi^2 - 2[-Log LiklihoodM4-(-Log Liklihood M3)]
Number of observations	747	747	
LR Chi^2	40.81	35.95	
df	10	7	
Prob > Chi^2	<0.001	<0.001	
Pseudo R^2	0.04	0.04	
Log Liklihood	-460.4246	-462.855	4.8603
Logistic Model for HIV, goodness of fit test			
	Model 3	**Model 4**	
Number of observations	747	747	
Number of covariate patterns	42	12	
Pearson Chi^2	25.73	1.32	
Prob > Chi^2	0.73	0.86	
Liklihood ratio test			
Assumption	**Model 4 is nested in Model 3**		
LR Chi^2	4.86		
df	3		
Prob > Chi^2	0.18		

Table 7 illustrates the comparison of model 3 and 4. The change in deviance between model and 3 and 4 was 4.8603 with 3 d.f. Thus there is no evidence (p=0.18) that model 3 gives a better explanation of the data than model 4, thus model 3 does not give an improvement over model 4, so we prefer the simpler model 4.

Table 8 Comparison of Model 2 and 6			
	Model 2	**Model 6**	**Difference in deviance**
			Chi^2 - 2[-Log LiklihoodM6-(-Log Liklihood M2)]
Number of observations	805	805	
LR Chi^2	48.52	42.82	
df	12	10	
Prob > Chi^2	<0.001	<0.001	
Pseudo R^2	0.05	0.04	
Log Liklihood	-489.2805	-492.1289	5.69686
Logistic Model for HIV, goodness of fit test			
	Model 2	**Model 6**	
Number of observations	805	805	
Number of covariate patterns	50	29	
Pearson Chi^2	31.81	20.05	
Prob > Chi^2	0.71	0.33	
Liklihood ratio test			
Assumption	**Model 6 is nested in Model 2**		
LR Chi^2	5.7		
df	2		
Prob > Chi^2	0.06		

Table 8 illustrates the comparison of model 2 and 6. The change in deviance between model and 2 and 6 was 5.69686; after conducting the likelihood ratio test, prob > chi^2 = 0.06 (3df). There is some evidence that model 2 is superior to model 6 (p=0.06); so we will interpret model 2 as it may give a richer interpretation than model 6.

We present parameter estimates for the models chosen above although we made further comparisons between models with the same observations but this is not presented. No comparison was made between model 6 and 3 or model 6 and 4 due to differing number of observations. Model 6 and model 4 and model 2 were therefore chosen as the ‘best’ models to fit the data. We also note that there are missing values in certain variables which results in three models, as we cannot compare these three models. One of the difficulties was the patterns of the missing data which could have been addressed using imputation methods however this was beyond of the scope of the analysis methodology of the assignment.

Table 9a. Results of Logistic Regression Model 2

Factors		Odds Ratio	P Value	95% CI
Sex				
	Male	1	reference	
	Female	3.22	<0.001	1.68-6.19
Site				
	Embalenhle	1	reference	
	Dunusa	0.69	0.239	0.38-1.28
Agegroup (yrs)				
	18-24	1	reference	
	25-34	2.59	0.003	1.37-4.87
	35-44	2.14	0.03	1.09-4.23
	45-50	1.38	0.53	0.50-3.80
STI				
Neisseria gonorrhoeae				
	Neg	1	reference	
	Pos	3.18	0.04	1.03-9.93
Chlamydia trachomatis				
	Neg	1	reference	
	Pos	1.59	0.1	0.91-2.78
Sex by agegroup interactions				
	Female (25-34)	0.41	0.03	0.18-0.90
	Female (35-44)	0.2	<0.001	0.08-0.47
	Female (45-50)	0.26	0.06	0.06-1.04
Site by sex interactions				
	Females (Dunusa)	2.31	0.04	1.03-5.19
Neisseria gonorrhoeae **by sex interactions**				
	Males (NG Positive)	0.32	0.108	0.08-1.29

Table 9a describes the output of the regression estimates for model 2. Further interpretations of interactions involving effects are based on the table of predicted probabilities (tables 9b, 9c and 9d).

Table 9b. Adjusting by Agegroup and STIs to determine the predicted probability of HIV by Site and Sex

Model 2	Male	95% CI	Female	95% CI
Community	29.73	24.49; 35.56	34.65	29.88; 39.74
Clinic	21.59	13.71; 32.32	46	34.35; 58.11

In table 9b we can see a strong sex by site interaction; in the community, the difference in HIV prevalence between females and males is much smaller than the difference in the clinic.

Table 9c. Adjusting by Site and STIs to determine the predicted probability of HIV by Agegroup and Sex - Model 2

Agegroup (yrs)	Male	95% CI	Female	95% CI
18-24	17.29	11.04; 26.02	43.68	34.83; 52.95
25-34	34.89	27.15; 43.52	44.89	37.76; 52.25
35-44	31.43	22.87; 41.48	24.58	17.72; 33.02
45-50	22.85	11.02; 41.46	21.49	10.00; 40.29

We observe a strong sex by age interaction (table 9c), which is consistent with previous findings. For females, the prevalence is already high in the 18-24 year old group, and then declines with increasing age. For males, the prevalence is lowest in the 18-24 old group, then peaks in the 25-34 year old group, remaining high in the 35-44 year group before declining in the 45-50 year old group.

Table 9d. Predicted probability of HIV by N. Gonorrhoeae and Sex				
Model 2	**Male**	**95% CI**	**Female**	**95% CI**
N. Gonorrhoeae				
Neg	26.69	22.13; 31.82	36.7	32.08; 41.58
Pos	51.96	26.50; 76.44	36.85	20.12; 57.48

Table 9d shows a strong sex by *N. Gonorrhoeae* interaction. The difference in HIV prevalence between males and females for having *N. Gonorrhoeae* is high and the odds are 3.18 (p=0.04) (table 9a). Males had a higher prevalence than females for *N. Gonorrhoeae* positives, while for *N. Gonorrhoeae* negatives, males has a lower HIV prevalence than females.

Table 10a. Results of Logistic Regression Model 4

Factors		Odds Ratio	P Value	95% CI
Sex				
	Male	1	reference	
	Female	3.06	0.001	1.61-5.80
Site				
	Embalenhle	1	reference	
	Dunusa	0.66	0.2	0.35-1.24
Agegroup (yrs)				
	18-24	1	reference	
	25-34	2.56	0.003	1.37-4.79
	35-44	2.02	0.04	1.03-3.94
	45-50	1.45	0.47	0.54-3.91
Sex and agegroup interactions				
	Female (25-34)	0.39	0.02	0.18-0.87
	Female (35-44)	0.2	<0.001	0.08-0.48
	Female (45-50)	0.23	0.04	0.05-0.92
Site and sex interactions				
	Females (Dunusa)	2.36	0.04	1.03-5.41

Table 10a describes the output of the regression estimates for model 4. All the main effects in the model, namely sex, site and agegroup, are also involved in interactions, so the interpretation of results from model 4 is based on the table of predicted probabilities.

Table 10b. Adjusting by Agegroup and STIs to determine the predicted probability of HIV by Site and Sex

Model 4	Male	95% CI	Female	95% CI
Community	29.46	24.04; 35.52	36.75	31.77; 42.02
Clinic	21.59	13.43; 32.83	44.82	33.03; 57.20

In table 10b we see a strong sex by age interaction, with females more likely to be HIV positive than males in the community, the difference in HIV prevalence

between females and males is much smaller than the difference in the clinic. The findings are similar to model 2.

Table 10c. Adjusting by Site and STIs to determine the predicted probability of HIV by Agegroup and Sex - Model 4

Agegroup (yrs)	Male	95% CI	Female	95% CI
18-24	18.23	11.79;27.10	42.86	34.14;52.02
25-34	36.35	28.51;44.98	44.96	37.84;52.29
35-44	30.99	22.53;40.92	24.13	17.39;32.46
45-50	18.22	11.79;27.10	43.21	34.49;52.37

There is a strong sex by age interaction in table 10c, which is consistent with previous findings (model 2). For females, the prevalence is already high in the 18-24 year old group, and then declines with increasing age. For males, the prevalence is lowest in the 18-24 old age group, and then peaks in the 25-34 year old group, drops in the 35-44 year group before declining rapidly in the 45-50 year old age group.

Table 11a. Results of Logistic Regression Model 6				
Factors		**Odds Ratio**	**P Value**	**95% CI**
Sex				
	Male	1	reference	
	Female	3.04	0.001	1.60-5.78
Site				
	Embalenhle	1	reference	
	Dunusa	0.7	0.25	0.39-1.28
Agegroup (yrs)				
	18-24	1	reference	
	25-34	2.56	0.003	1.37-4.87
	35-44	2.05	0.04	1.09-4.23
	45-50	1.42	0.49	0.50-3.80
STI GPCR				
	Neg	1	reference	
	Pos	1.73	0.11	0.88-3.40
Sex and agegroup interactions				
	Female (25-34)	0.4	0.02	0.18-0.88
	Female (35-44)	0.19	<0.001	0.08-0.47
	Female (45-50)	0.24	0.04	0.06-0.95
Site and sex interactions				
	Females (Dunusa)	2.3	0.04	1.03-5.11

Table 11a describes the output of the regression estimates for model 6. Participants with *N. Gonorrhoeae* were 1.73 times more likely to be HIV positive, {(95% CI, 0.88; 3.40 p=0.11)}. This effect is not significant however shows that those infected with *N. Gonorrhoeae* are at an increased risk of HIV. Further interpretations of interactions are based on the table of predicted probabilities (tables 11b and 11c).

Table 11b. Adjusting by Agegroup and STIs to determine the predicted probability of HIV by Site and Sex

Model 6	Male	95% CI	Female	95% CI
Community	29.88	24.69; 35.64	34.91	30.15; 39.99
Clinic	23	14.83; 33.89	44.09	32.64; 56.19

In table 11b we see a strong sex by site interaction; in the community, the difference in HIV prevalence between females and males is much smaller than the difference in the clinic. The findings are similar to model 2 and 4.

Table 11c. Adjusting by Site and STIs to determine the predicted probability of HIV by Agegroup and Sex - Model 6

Agegroup (yrs)	Male	95% CI	Female	95% CI
18-24	18.22	11.78; 27.11	42.59	33.88; 51.79
25-34	36.29	28.45; 44.93	44.74	37.61; 52.09
35-44	31.35	22.85; 41.33	24	17.27; 32.36
45-50	24.08	11.88; 42.72	20.38	9.41; 38.68

There is also a strong sex by age interaction in table 8c, which is consistent with previous findings (model 2 and 4). For females, the prevalence is already high in the 18-24 year old group, and then declines with increasing age. For males, the prevalence is lowest in the 18-24 old age group, then peaks in the 25-34 year old age group, drops in the 35-44 year group before declining rapidly in the 45-50 year old group.

4. Discussion

The study is consistent with the fact that HIV is a public health challenge in South Africa. HIV seroprevalence among study participants was high with a crude study seroprevalence of 33.72% (95% CI 30.54; 37.02).

Community STI seroprevalence was high, 11.14% (95% CI 9.05; 13.51); with *N. gonorrhoeae* being at 4.77% (95% CI 3.41; 6.47) and *C. trachomatis* being at 7.79% (95% CI 6.04; 9.87), with no significant difference between males and females.

Having any STI was a risk factor for HIV {unadjusted OR=1.76, (95% CI, 1.12; 2.75, p=0.01)}. Females were at higher risk of acquiring HIV than males {OR=1.51, (95% CI, 1.12; 2.01, p=0.006). Both *N. gonorrhoeae and C. trachomatis* were risk factors for HIV infection (results not shown).

The distribution of agegroup by sex showed the majority of females in the study were between the ages of 21 and 35 years and peaking in the 21-25 year agegroup. The distribution of males was somewhat more uniform across the agegroups and the distribution was bimodal when stratified by site.

The findings of the regression estimates and predicted probabilities together with the descriptive statistics suggest that the distribution of HIV in this setting is complex as shown by the predicted probabilities from the three selected logistic regression models.

All models gave similar predicted HIV seroprevalences in the two-way breakdown by site and sex, as described in the results,

We notice from the predicted probability output that HIV seroprevalence for females is 34.65% and 29.73% for males at the community level, table 9b, model 2. The difference between female and male was 4.92% whilst the difference in HIV seroprevalence between sexes at the clinic level was 24.41% for the same model. Similar findings are noted in model 4. Thus the degree of difference in HIV prevalence between sexes differs significantly between the two sites; this is interesting and shows that the comparison of seroprevalence between males and females might differ in different settings, thus further investigation into this is required.

This distribution of HIV across age strata from the predicted probability output is more consistent with the current epidemiology of HIV in South Africa (10, 15).

The effect of sex differed for site; for females the HIV seroprevalence was higher in the health care facility, while for males the HIV prevalence was higher in the community. This is interesting and may be explained or accounted for by the different health seeking profiles of males and females.

Nicolosi et al. compared the efficiency of male to female and female to male sexual transmission of HIV (24). Using a logistic regression analysis and controlling for multiple confounders, it was found that the efficiency of male to female transmission was 2.3 times greater than female to male transmission. It is suggested that between gender differences in the contact surfaces and intensity of exposure during sexual intercourse are possible reasons for the gender differential.

Gregson et. al report that substantial age differentials between female and male sexual partners in Manicaland are the major behavioural determinant of the more rapid rise in HIV prevalence in young women than in men (25). After controlling for confounders, it was demonstrated that a large gender effect remains after controlling for confounders {OR = 6.04 ((95% CI. 1.49; 24.47)}.

Glynn et. al. suggest that behavioural factors could not fully explain the discrepancy in HIV prevalence between males and females (26). Despite the tendency for women to have older partners, young men were at least as likely to encounter an HIV infected partner as young women. It is likely that the greater susceptibility of women to HIV infection is an important factor both in explaining the male to female discrepancy in HIV prevalence and in driving the epidemic.

The results of this study could point to gender specific health seeking behaviour patterns not only for HIV but for general clinic disease showing us, as documented in other studies, that females have different health seeking behaviour patterns than males (11, 16).

4.1 Study Limitations

The sampling frame of the initial study was not available for the purposes of reporting and informing the secondary analysis of this study. The initial design of the seroprevalence survey was not made available and thus the participant response rates are not known. We are therefore unable to quantify the non-response rates or mention the bias in sampling. The availability of this information would better inform us of the nature of the health seeking behaviour pattern of the population under investigation. Also we must take into account the unequal distribution of the sample between sites; the secondary analysis is not able to comment on the methodology of the initial design.

Furthermore, the secondary analysis included only disease prevalence and 3 demographic variables into the descriptive and regression models, as these were the only variables made available to us. The lack of socio-demographic variables proves to be a serious limitation in the analysis as this would have been better able to inform us about other risk factors in this community during the period of the study.

Due to the fact that the binary regression models are nonlinear, no single approach to interpretation can fully describe the relationship between a variable and the outcome. In general, the estimated parameters from the binary regression model do not provide useful information for understanding the relationship between independent variables and the outcome. With the exception of the rarely used method of interpreting the latent variable, substantive meaningful interpretations are based in the predicted probabilities and functions of those probabilities (23).

It is also clear from this study that it may not be appropriate to compare HIV prevalence from hospital based studies and community studies as huge site and sex as well as site and age group interactions are often encountered in addition to being documented in the secondary analysis of this assignment. The analysis would have been more manageable if analysis was analysed separately by site resulting in fewer levels of interaction to manage however the study highlights crucially the difficulty of combining data from the 2 sites.

Also note that this was a cross sectional study; no mention of clustering was made available and could have impacted differently on the analysis. Cross sectional studies often deal with exposures that cannot change; for such exposures current information such as in this study is useful however for variable exposure, current information is less desirable (27). Fortunately this was not the case in this study. This study had no etiological objectives and no reference to cause of the primary outcome of disease was made.

We should not use this study to make generalizations on the nature and spread of the HIV epidemic nationally or even in Mpumalanga province, as the initial study was designed to measure the burden of disease in a mining community, of which the population may be at a higher risk for contraction of infectious disease. The risk profile of this community differs considerably from the general population however we must take into account that studies like this are required on a larger

National Scale in order to ascertain risk profiles of communities so that interventions to limit the spread of HIV or manage the care of HIV infected patients can be better planned.

5. Conclusion and Recommendation

This study informs us that there is a need for more information than what is provided by the National Antenatal survey. With the advent of community surveys such as the Nelson Mandela HSRC study, we will be able to better understand the epidemic at the community level. At the current stage of this epidemic, community studies will be better able to inform public health professionals on strategies for planning and operationalizing the current comprehensive care management and treatment program.

Differences in the spread of HIV in the male and female population can be accounted for by a complex interplay of sexual behaviour and biological factors that affect the probability of HIV transmission per sex act (28). Sexual behaviour patterns are determined by cultural and socioeconomic contexts and in sub-Saharan Africa, some traditions and socioeconomic developments have contributed to the extensive spread of HIV infection (28). Resulting from this spread of HIV, we are faced with challenges in quantifying differential infection rates in population groups thus making planning for targeted interventions difficult. The subordinate position of women, impoverishment and decline of social services, rapid urbanization and modernization are factors that contribute to the evolutionary epidemiological profile and public health interventions are challenged to address this.

The National Strategic Plan (29) clearly states that there is a need to scale up the management of HIV care and treatment programs in South Africa, therefore more studies focused on describing the epidemiology of HIV for populations at high and low risk is required. As gathered from this survey, the difference in the prevalence of HIV as seen from the predicted probabilities of HIV across males and females calls for further investigation on a broader scale and therefore we should be encouraging large community based surveys such as the HSRC survey to continue and better inform the management of the National

Comprehensive Care Management and Treatment program (30).

6. References

1. Grosskurth H, Gray RG, Hayes RJ, et al. Control of sexually transmitted diseases for HIV-1 prevention: Understanding the implications of the Mwanza and Rakai Trials. Lancet 2000; 255: 1991-1987
2. World Health Organization/WPRO-SEARO. HIV/AIDS in Asia and the Pacific Region, 2001
3. UNAIDS/WHO AIDS epidemic update, Global summary: December 2005
4. UNAIDS/WHO AIDS epidemic update, sub-Saharan Africa: December 2005
5. Buve A, Carael M, Hayes R, Robinson NJ. Variations in HIV prevalence between urban areas in sub-Saharan Africa: do we understand them? AIDS 1995; (Suppl A): S103-S109
6. Auvert B, Taljaard D, Lagarde E, Sobngwi-Tambekou J, Sitta R, et al. (2005) Randomized, controlled intervention trial of male circumcision for reduction of HIV infection risk: The ANRS 1265 trial.PLoS Med 2(11): e298
7. Steven J Reynolds, Mary E Shepherd, Arun R Risbud, Raman R Gangakhedkar, Ronald S Brookmeyer, Anand D Divekar, Sanjay M Mehendale, Robert C Bollinger. Male circumcision and risk of HIV-1 and other sexually transmitted infections in India. *Lancet* 2004; 363: 1039–1040
8. Carael M, Holmes K. Dynamics of HIV epidemics in sub-Saharan Africa: introduction AIDS 2001; 15 (Suppl 4): S1-S4
9. Buve A, Carael M, Hayes R, et al. Multicentre study on factors determining differences in rates of spread of HIV in sub-Saharan Africa: methods and general population prevalence of HIV infection. AIDS 2001 15 (Suppl): S5-S14

10. Boisson E, Nicoll A, Zaba B, et. al. Interpreting HIV seroprevalence data from pregnant women. Journal of Acquired Immune Deficiency Syndrome and Human Retrovirology, vol 13, pp. 434-439, 1996.
11. Shisana O, Simbayi L. Nelson Mandela/HSRC Study of HIV/AIDS, South African National HIV Prevalence, Behavioral Risks and Mass Media, Household Survey 2002.
12. Department of Health, 2006. National HIV and syphilis antenatal seroprevalence survey in South Africa 2005.
13. Department of Health, 2007. National HIV and syphilis antenatal seroprevalence survey in South Africa 2006.
14. Department of Health, 2005. National HIV and syphilis antenatal seroprevalence survey in South Africa 2004.
15. Department of Health 2002b. National HIV and syphilis seroprevalence survey if women attending public antenatal clinics in South Africa 2001.South Africa
16. HIV Prevalence, Incidence, Behavior and Communication Survey 2005, by collaborative research of Human Science Research Council (HSRC), Medical Research Council (MRC) and Centre for AIDS Development, Research and Evaluation (CADRE), of South Africa.
17. 1997 Epidemiological Comments, Department of Health.
18. Pers. Comm. Basia Zaba (Urassa M, Zaba B, Boerma T, Schenk K. Indirect estimated of Male HIV prevalence from Ante-natal clinic surveillance.)
19. Council for Scientific and Industrial Research, HIV seroprevalence rates in Embalenhle and Dunusa, Mpumalanga, South Africa (unpublished report).
20. Hamilton, C.H. Statistics with STATA, pages 272-273
21. Dupont, W.D.; Statistical Modeling for Biomedical Researchers, A Simple Introduction to the Analysis of Complex Data, pages 116-118.
22. Betty R. Kirkwood and Jonathan A.C. Sterne, Medical statistics 2nd edition, Chapter 28 Statistical Modelling, pages 309-314

23. Long J.S.; Freese J. Regression models for categorical dependant variables using STATA, pages 131-134
24. Nicolosi A, Leite MLC, Musicco M, et al. The efficiency of male to female and female to male sexual transmission of the Human Immunodeficiency Virus: A study of 730 stable couples. Epidemiology, 1994 Nov; 5(6):565-570
25. Gregson S, Nyamukaoa C, Garnett G, et al. Sexual mixing patterns and sex-differentials in teenage exposure to HIV infection in rural Zimbabwe. The Lancet, Volume 359, issue 9321, pages 1896-1903
26. Glyn J.R, Carael M, Auvert, B, et al. Why do young women have a much higher prevalence of HIV than young men? A study in Kismu, Kenya and Ndola, Zambia. AIDS: Volume 15 Supplement 4 August 2001 pp S51-S60
27. Rothman, K.J., Greenland, S.; Modern Epidemiology, 2nd Edition, pages 75-76
28. Buve A, Bishikwabo-Nsarhaza K, Mutangadura G. The spread and effect of HIV-1 infection in sub-Saharan Africa. The Lancet, Volume 359, issue 0322, pages 2011-2007
29. HIV and AIDS and STI Strategic Plan for South Africa, 2007-2011, Department of Health, 2007
30. Operational plan for comprehensive HIV and AIDS Care, Management and Treatment for South Africa, 19 November 2003